Covid-Global Data Analysis

Sarbani Banerjee

19/08/2020

Introduction

The novel human coronavirus disease COVID-19 has become the fifth documented pandemic since the 1918 flu pandemic. COVID-19 was first reported in Wuhan, China, and subsequently spread worldwide.The earliest date of symptom onset was 1 December 2019 and on 12 February 2020 the disease officially named as COVID-19. Since then more than 6 months have passed and the whole world is fighting against this deadly disease.In this report we will try to analyse the scenario of the different countries of the world in this last six months. The data has been taken from Kaggle.Though the data has been updated daily, we will only take the time span from 12 February to 12 August 2020 counting 6 months.

```
covid<- read.csv("covid_six.csv")
head(covid)
```

```
##         Date Country_Region               Province_State positive active
## 1 12-02-2020      Australia                   All States       15     NA
## 2 12-02-2020      Australia Australian Capital Territory        0     NA
## 3 12-02-2020      Australia              New South Wales        4     NA
## 4 12-02-2020      Australia           Northern Territory        0     NA
## 5 12-02-2020      Australia                   Queensland        5     NA
## 6 12-02-2020      Australia              South Australia        2     NA
##   hospitalized hospitalizedCurr recovered death total_tested daily_tested
## 1           NA               NA        NA    NA           NA           NA
## 2           NA               NA        NA    NA           NA           NA
## 3           NA               NA        NA    NA           NA           NA
## 4           NA               NA        NA    NA           NA           NA
## 5           NA               NA        NA    NA           NA           NA
## 6           NA               NA        NA    NA           NA           NA
##   daily_positive
## 1              0
## 2              0
## 3              0
## 4              0
## 5              0
## 6              0
```

```
tail(covid)
```

```
##              Date Country_Region Province_State positive active
hospitalized
## 18211 12-08-2020  United States       Virginia   102521  86922
14528
```

```
## 18212 12-08-2020  United States     Washington    64151      NA
6102
## 18213 12-08-2020  United States  West Virginia     8008   1895
NA
## 18214 12-08-2020  United States      Wisconsin    66654  13286
5125
## 18215 12-08-2020  United States        Wyoming     3086    480
187
## 18216 12-08-2020        Vietnam     All States       NA      NA
NA
##       hospitalizedCurr recovered death total_tested daily_tested
daily_positive
## 18211             1281     13247  2352      1287556        16011
776
## 18212              383        NA  1716      1014258          504
504
## 18213              135      5960   153       335239         4630
133
## 18214              364     52350  1018      1090377         9977
531
## 18215               15      2577    29        59331          479
13
## 18216               NA        NA    NA       621823        51545
NA
```

```
dim(covid)
```

```
## [1] 18216    12
```

```
lapply(covid,class)
```

```
## $Date
## [1] "character"
##
## $Country_Region
## [1] "character"
##
## $Province_State
## [1] "character"
##
## $positive
## [1] "integer"
##
## $active
## [1] "integer"
##
## $hospitalized
## [1] "integer"
##
## $hospitalizedCurr
## [1] "integer"
```

```
##
## $recovered
## [1] "integer"
##
## $death
## [1] "integer"
##
## $total_tested
## [1] "numeric"
##
## $daily_tested
## [1] "integer"
##
## $daily_positive
## [1] "integer"
```

From the column Province_State we can see that different levels i.e country/ state has been given. Since we are analysing the data at country level we need to filter this out for All States.

```
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

covid_allstate<- filter(covid,Province_State=="All States")
head(covid_allstate)

##         Date Country_Region Province_State positive active hospitalized
## 1 12-02-2020      Australia     All States       15     NA           NA
## 2 12-02-2020        Czechia     All States        0     NA           NA
## 3 12-02-2020         Israel     All States       14     NA           NA
## 4 12-02-2020         Russia     All States        2     NA           NA
## 5 12-02-2020 United Kingdom     All States       15     NA           NA
## 6 12-02-2020  United States     All States       18     NA           NA
##   hospitalizedCurr recovered death total_tested daily_tested
daily_positive
## 1               NA        NA    NA           NA           NA
0
## 2               NA        NA    NA           75            1
0
## 3                0        NA     0          306           35
0
```

```
## 4               NA         2    NA           NA           NA
0
## 5               NA        NA    NA         1758          400
1
## 6               NA        NA     0           18            1
1

dim(covid_allstate)

## [1] 5577   12
```

Which countries have had the highest number of deaths due to COVID-19? To answer this question we need to group the data country wise and calculate the maximum death from death column. But since there are presence of missing values in the death column we will replace NA values by zero

```
covid_allstate_death <- covid_allstate %>%
  select(Country_Region,death)
covid_allstate_death[is.na(covid_allstate_death)]=0
covid_death<- covid_allstate_death %>%
  group_by(Country_Region) %>%
  summarise(total_death=max(death)) %>%
  arrange(-total_death)

## `summarise()` ungrouping output (override with `.groups` argument)

covid_death

## # A tibble: 134 x 2
##    Country_Region total_death
##    <chr>                <dbl>
##  1 United States       157776
##  2 Italy                35225
##  3 United Kingdom       33186
##  4 Canada                9006
##  5 Belgium               8903
##  6 Sweden                5690
##  7 Turkey                5458
##  8 Russia                5215
##  9 Bangladesh            3513
## 10 Poland                1359
## # ... with 124 more rows
```

The result shows that USA,UK,Italy have suffered from high mortality during this six months.But due to lack of availability of data for other countries we could not get the complete picture .

Which countries have had the highest number of positive cases against the number of tests?

To answer this question we need the Country_Region ,daily_tested, daily_positive columns. Since the data is updated daily, there are missing values in case of unavailability of data. We will replace those missing values with the averaged across number of days in between. Again we need to group the data and sum it up and calculate the ratio.

```
covid_tp<- covid_allstate %>%
  select(Country_Region,daily_tested,daily_positive)

new_tp_rev <- covid_tp %>%  group_by(Country_Region) %>%
  mutate_all(funs(ifelse(is.na(.), round(mean(., na.rm = TRUE)),.)))
```

```
## `mutate_all()` ignored the following grouping variables:
## Column `Country_Region`
## Use `mutate_at(df, vars(-group_cols()), myoperation)` to silence the
message.
```

```
## Warning: `funs()` is deprecated as of dplyr 0.8.0.
## Please use a list of either functions or lambdas:
##
##   # Simple named list:
##   list(mean = mean, median = median)
##
##   # Auto named with `tibble::lst()`:
##   tibble::lst(mean, median)
##
##   # Using lambdas
##   list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.
```

```
covid_tp_sum<- new_tp_rev %>% group_by(Country_Region) %>%
  summarise(tested_sum=sum(daily_tested),tested_positive=sum(daily_positive))
%>%
  arrange(-tested_sum)
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
covid_tp_sum<- head(covid_tp_sum,10)

covid_tp_percentage<- covid_tp_sum %>%
  mutate(percentage=((tested_positive/tested_sum)*100)) %>%
  arrange(-percentage)
covid_tp_percentage
```

```
## # A tibble: 10 x 4
##    Country_Region tested_sum tested_positive percentage
##    <chr>               <dbl>           <dbl>      <dbl>
##  1 Peru              2251380          227372      10.1 
##  2 United States    67299150         5527882       8.21
##  3 Turkey            4423138          219500       4.96
##  4 Israel            1966661           85056       4.32
##  5 Russia           12397659          432269      3.49
##  6 Italy             7408577          252963      3.41
##  7 Canada            4652621          122950      2.64
##  8 Brazil            3099118           33025      1.07
##  9 India            23862244          206557      0.866
## 10 Australia         6482803           22112      0.341
```

After summing up the daily tested and daily positive results we take the top 10 countries where most numbers of test has been conducted.USA,India,Russia ,Italy are the top 4 countries where highest number of test have been conducted. Then we calculated the percentage of people tested positive out of total tested, it will give us an idea the extent of spread of the disease among the population.Strikingly Peru has the highest percentage of positive cases out of top 10 tested countries followed by USA. Though India has tested a large number of population the percentage of positive cases is comparetively lower.

It is clear that in this fight against the virus, each country has defended itself as best it can. We want to quantify this effort for the top ten tested cases countries.at the population level. **Which countries have made the best effort in terms of the number of tests conducted related to their population?**

```
population<- c(331002651, 1380004385, 145934462 , 60461826,
25499884,37742154,84339067,212559417,32971854, 8655535)
covid_tp_sum<- data.frame(covid_tp_sum,population)
covid_population<- covid_tp_sum %>%
  mutate(percentage_tested=(tested_sum/population)*100,
         percentage_positive=(tested_positive/population)*100) %>%
select(Country_Region,percentage_tested,percentage_positive) %>% arrange(-
percentage_tested)
covid_population
```

```
##    Country_Region percentage_tested percentage_positive
## 1       Australia         25.422873          0.08671412
## 2          Israel         22.721426          0.98267756
## 3   United States         20.331907          1.67004161
## 4          Canada         12.327386          0.32576307
## 5           Italy         12.253313          0.41838465
## 6          Russia          8.495361          0.29620762
## 7            Peru          6.828187          0.68959422
## 8          Turkey          5.244471          0.26025899
## 9           India          1.729143          0.01496785
## 10         Brazil          1.458001          0.01553683
```

The data states that though countries like India, Russia , Italy have conducted large number of tests, but in comparison to total population of these countries the number is not very high.In fact for developing nations like India,Brazil this percentage is quite low. To take efficient measures towards defeating the disease the developing nations need to enhance their infrastructure and increase their pace of testing.

Now we will do a comparative study between India (developing nation) & USA (developed nation) to understand their testing pattern during these six months

```r
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```r
covid_allstate$DateR <- dmy(covid_allstate$Date)
covid_allstate$month<- month(covid_allstate$DateR)
covid_India<- filter(covid_allstate,Country_Region=="India")
covid_India<-covid_India %>% select(month,daily_tested)
covid_USA<- filter(covid_allstate,Country_Region=="United States")
covid_USA<- covid_USA %>% select(month,daily_tested)
covid_India <- covid_India %>%  group_by(month) %>%
  mutate_all(funs(ifelse(is.na(.), round(mean(., na.rm = TRUE)),.)))
```

```
## `mutate_all()` ignored the following grouping variables:
## Column `month`
## Use `mutate_at(df, vars(-group_cols()), myoperation)` to silence the
message.
```

```r
covid_USA <- covid_USA %>%  group_by(month) %>%
  mutate_all(funs(ifelse(is.na(.), round(mean(., na.rm = TRUE)),.)))
```

```
## `mutate_all()` ignored the following grouping variables:
## Column `month`
## Use `mutate_at(df, vars(-group_cols()), myoperation)` to silence the
message.
```

```r
covid_India<- covid_India %>%
  group_by(month) %>%
  summarise(total_test_India=sum(daily_tested,na.rm = TRUE))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```r
covid_USA<- covid_USA %>%
  group_by(month) %>%
  summarise(total_test_USA=sum(daily_tested,na.rm = TRUE))
```

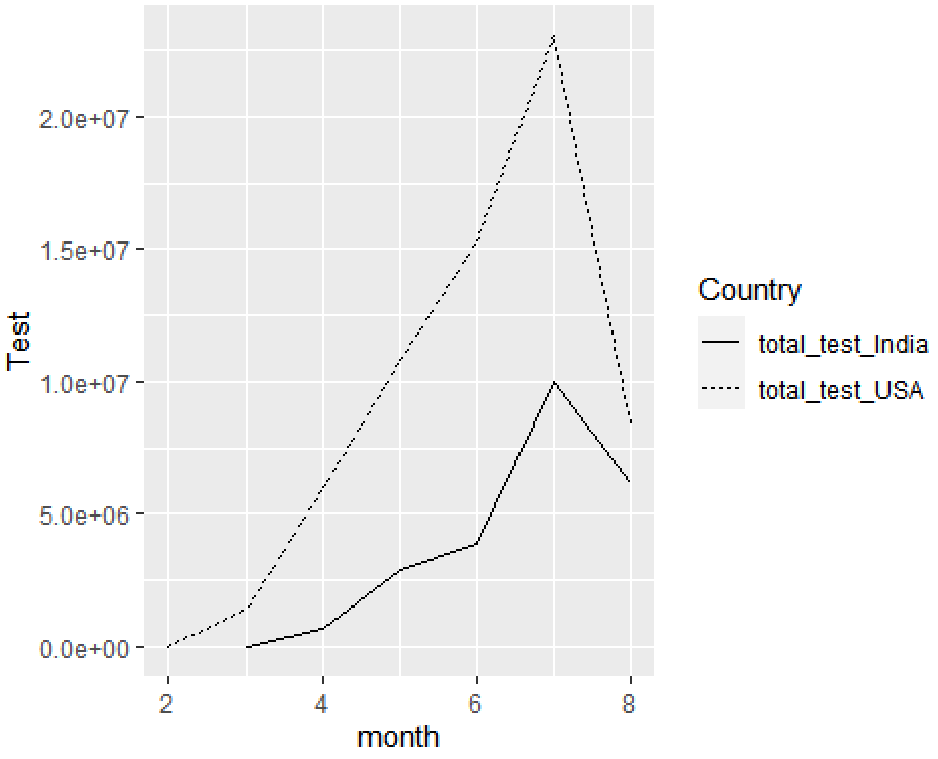
```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```r
covid_India_USA<- covid_USA %>% full_join(covid_India, by="month")
```

```r
library(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 4.0.2
```

```
##
## Attaching package: 'tidyr'

## The following object is masked _by_ '.GlobalEnv':
##
##     population

covid_India_USA_reshape<- covid_India_USA %>%
  pivot_longer(cols=c(total_test_USA,total_test_India),
               names_to="Country",values_to="Test")
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.0.2

ggplot(data = covid_India_USA_reshape) +
  aes(x = month, y = Test, lty = Country) +
  geom_line()

## Warning: Removed 1 row(s) containing missing values (geom_path).
```



As the result shows over the month in USA number of tests is much higher than India. Though for both the countries number of tests conducted are increasing over the months.